Réseaux

Haute disponibilité

Sommaire

Généralités

La haute disponibilité désigne le fait qu'un service ou une architecture matérielle possède un taux de disponibilité "correct"

Le taux de disponibilité se calcule en pourcentage d'unités de temps où le service est disponible.

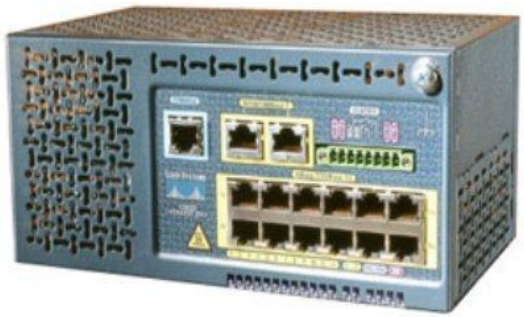
Il s'exprime souvent en années, par exemple:

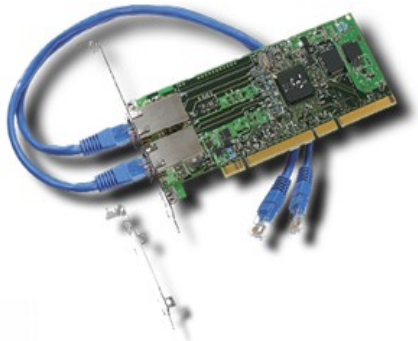
99%	→ 3,65 jours /an indisponible
99,99%	→ 4,38 heures / an indisponible
99,999%	→ ~ 5 minutes / an indisponible
99,9999%	→ ~ 30 secondes / an indisponible

Plusieurs techniques permettent d'améliorer la disponibilité, en fonction du niveau où l'on se situe:

au niveau matériel :

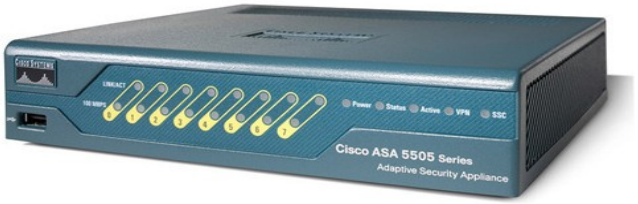
- pour les disques durs: le raid (d'un niveau 1, 5 ou 10), les SAN ;
- pour les cartes réseaux : teaming / bonding (802.3ad) ;
- pour les serveurs : les lames ;

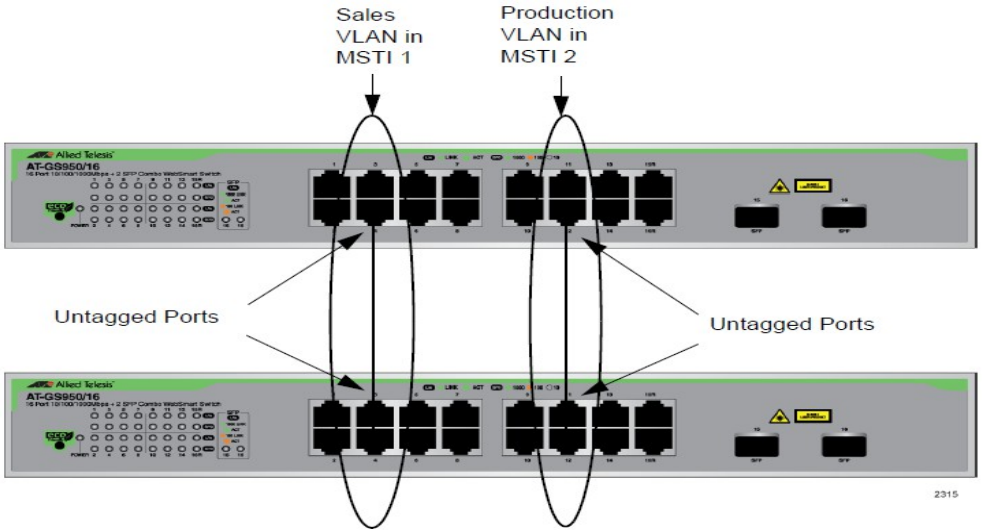




au niveau réseau :

- spanning tree (OSI 2) ;
- roaming (802.11r) ;
- routes statiques flottantes (OSI3) ;
- partage d'IP dynamique (HSRP).





au niveau service :

- système de maître / esclave ;
- système de partage d'IP (UCARP, VRRP, ...) ;
- système de synchronisation à chaud (DRDB) ;
- reverse proxy / balance de charge (Squid / LVS / HA Proxy)

au niveau datacenter :

- images d'exploitation / copie à chaud (virtualisation)
- gestion de configurations (puppet)
- constitution de noeuds (nodes)

Virtually Anything Goes

Redondance au niveau matériel

Les disques durs :

Le mécanisme le plus utilisé pour redonder l'information au niveau des disques dur est le RAID (Redundant Array of Independent Disk) en opposition au SLED (Single Large Expensive Disk).

Le RAID permet de mixer des disques durs "classiques" dans une matrice qui sera plus performante (plus d'espace, plus rapide, plus sécurisée)

Il existe deux types de RAID : matériels et logiciels

Les niveaux de RAID qui vont nous intéresser sont 0, 1, 5 et 10

Le RAID logiciel

Dans ce cas, le contrôle du **RAID** est **intégralement assuré** par une **couche logicielle** du système d'exploitation. Cette couche s'intercale entre la couche d'abstraction matérielle (pilote) et la couche du système de fichiers.

Avantages

- **peu cher** ;
- très **souple** (administration logicielle) ;
- la grappe est **compatible** avec toutes les machines utilisant le même OS.

Le RAID logiciel

Inconvénients

- la couche d'abstraction matérielle peut manquer de fonctions importantes comme la détection et le diagnostic des défauts matériels et/ou la prise en charge du remplacement à chaud (Hot-swap) des unités de stockage ;
- la gestion du RAID **monopolise des ressources systèmes** (CPU et bus système)
- l'utilisation du RAID sur le disque système n'est pas toujours possible.

Le RAID logiciel

Implémentations

- Sous Windows XP et + seulement le RAID 0 et 1 sont gérés et sous Windows Serveur le RAID 5 est supporté ;
- Sous MAC seulement le RAID 0 et 1 sont supportés ;
- Sous Linux (noyau 2.6) les RAID 0, 1, 4, 5, 6 et 10 sont supportés ainsi que les combinaisons de ces modes (ex. 0+1)

Le RAID matériel

Une carte ou un composant est dédié à la gestion des opérations, dotée d'un processeur spécifique, de mémoire, éventuellement d'une batterie de secours, et est capable de gérer tous les aspects du système de stockage RAID grâce au ***firmware***.

Avantages

- détection des défauts, remplacement à chaud des unités défectueuses, possibilité de reconstruire de manière transparente les disques défaillants ;
- charge système est allégée ;
- la vérification de cohérence, les diagnostics et les maintenances sont effectués en arrière-plan par le contrôleur sans solliciter de ressources système.

Le RAID matériel

Inconvénients

- les contrôleurs RAID matériels utilisent chacun leur propre système pour gérer les unités de stockage et donc aucune donnée ne pourra être récupérée si le contrôleur RAID n'est pas exactement le même (firmware compris) ;
- les cartes d'entrée de gamme possèdent des processeurs peu puissants et donc les performances sont moins bonnes ;
- l'entrée de gamme se situe aux alentours de 200€ mais les cartes plus performantes dépassent souvent les 1 000€.
- le contrôleur RAID est lui-même un composant matériel, qui peut tomber en panne ("***single-point-of-failure***") ;

Le RAID 0

Egalement connu sous le nom d'« entrelacement de disques » ou de « volume agrégé par bandes » (***stripping***), c'est une configuration permettant d'augmenter significativement les performances de la grappe en faisant travailler **N** disques durs en parallèle (avec **N > 2**).

- Capacité

La capacité totale est égale à celle du plus petit élément de la grappe multiplié par le nombre d'éléments présents dans la grappe.

L'espace excédentaire des autres éléments de la grappe restera inutilisé, il est donc conseillé d'**utiliser des disques de même capacité**.

Le RAID 0

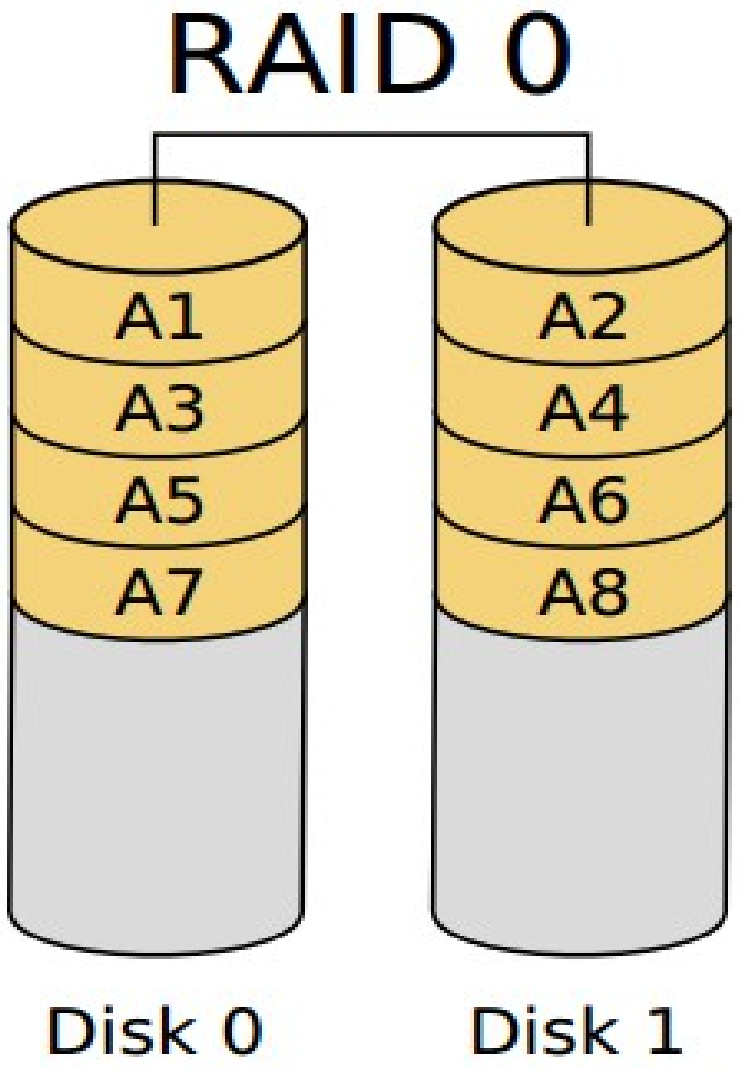
- Fiabilité

Le défaut de cette solution est que la perte d'un seul disque entraîne la perte de toutes ses données.

- Coût

Dans un RAID 0, qui n'apporte aucune redondance, tout l'espace disque disponible est utilisé (tant que tous les disques ont la même capacité).

Le RAID 0



Le RAID 1

RAID miroir (mirroring) consiste en l'utilisation de N disques redondants (avec N > 2), chaque disque de la grappe contenant à tout moment exactement les mêmes données.

- Capacité

La capacité totale est égale à celle du plus petit élément de la grappe. L'espace excédentaire des autres éléments de la grappe restera inutilisé. Il est donc conseillé d'utiliser des éléments identiques.

Le RAID 1

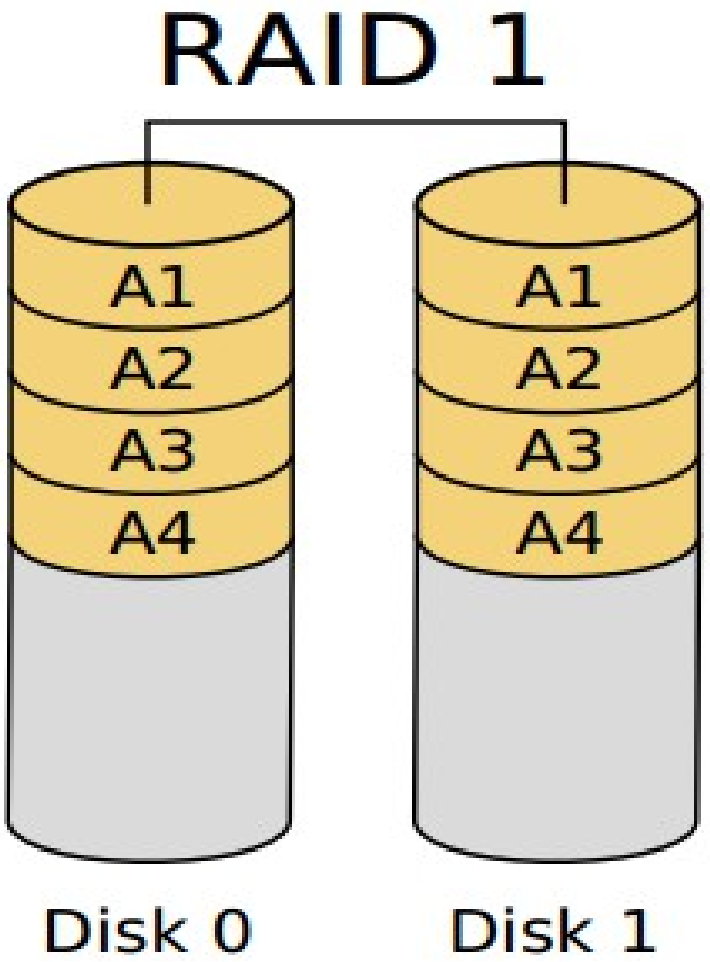
- Fiabilité

Cette solution offre un excellent niveau de protection des données. Elle accepte une défaillance de N-1 éléments.

- Coût

Les coûts de stockage sont élevés et directement proportionnels au nombre de miroirs utilisés alors que la capacité utile reste inchangée. Plus le nombre de miroirs est élevé, et plus la sécurité augmente, mais plus son coût devient prohibitif.

Le RAID 1



Le RAID 5

Le RAID 5 combine le stripping à une parité répartie pour un ensemble à redondance N+1.

La parité, qui est incluse avec chaque écriture, se retrouve répartie circulairement sur les différents disques.

Chaque bande est donc constituée de N blocs de données et d'un bloc de parité.

En cas de défaillance, on peut "retrouver" les données à partir des N-1 autres blocs de données et du bloc de parité.

Pour limiter le risque, il est courant de dédier un disque dit de "spare" qui en régime normal est inutilisé et en cas de panne prendra la place du disque défaillant.

Le RAID 5

- Capacité

La capacité totale est égale à (N-1) x C avec N le nombre de disque et C la capacité du plus petit élément de la grappe.

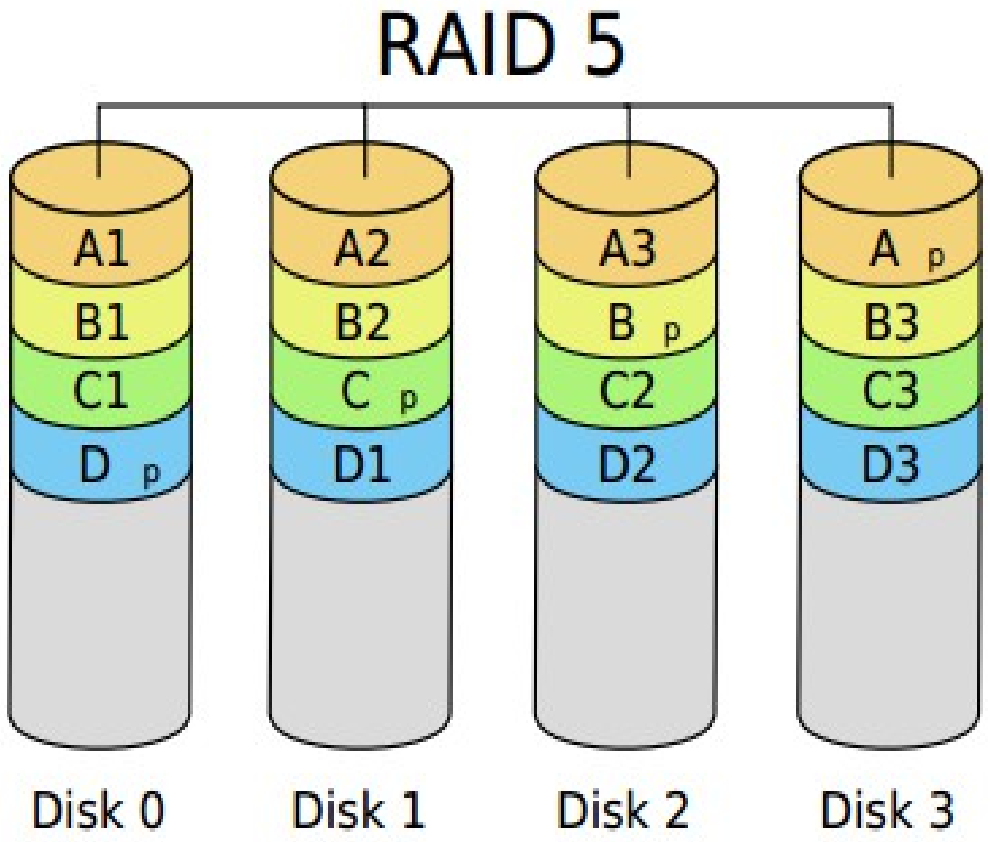
- Fiabilité

Cette solution offre un excellent niveau de protection des données. Elle accepte une défaillance de N-1éléments.

- Coût

Les coûts de stockage sont corrects et égaux à 1 disque en plus dans le meilleur des cas, à 1 + N disques de spare dans le pire des cas.

Le RAID 5


RAID 5
A1
B1
C1
D p
A2
B2
C p
D1
A3
B p
C2
D2
A p
B3
C3
D3
Disk 0
Disk 1
Disk 2
Disk 3

Les cartes réseaux

Il est possible d'agréger des liens grâce à la norme 802.3ad pour :

- Gagner en bande passante (agrégation de lien / channel bonding) ;
- Augmenter la tolérance de panne (teaming).

Protocole ouvert → Link Control Aggregation Protocol (LCAP)

Protocole propriètaire (CISCO) → Port Aggregation Protocol (PagP)

Limitation: les ports d'une agrégation doivent être de même nature cuivre / SX (multimode) / LX (monomode) et de même vitesse en full-duplex.

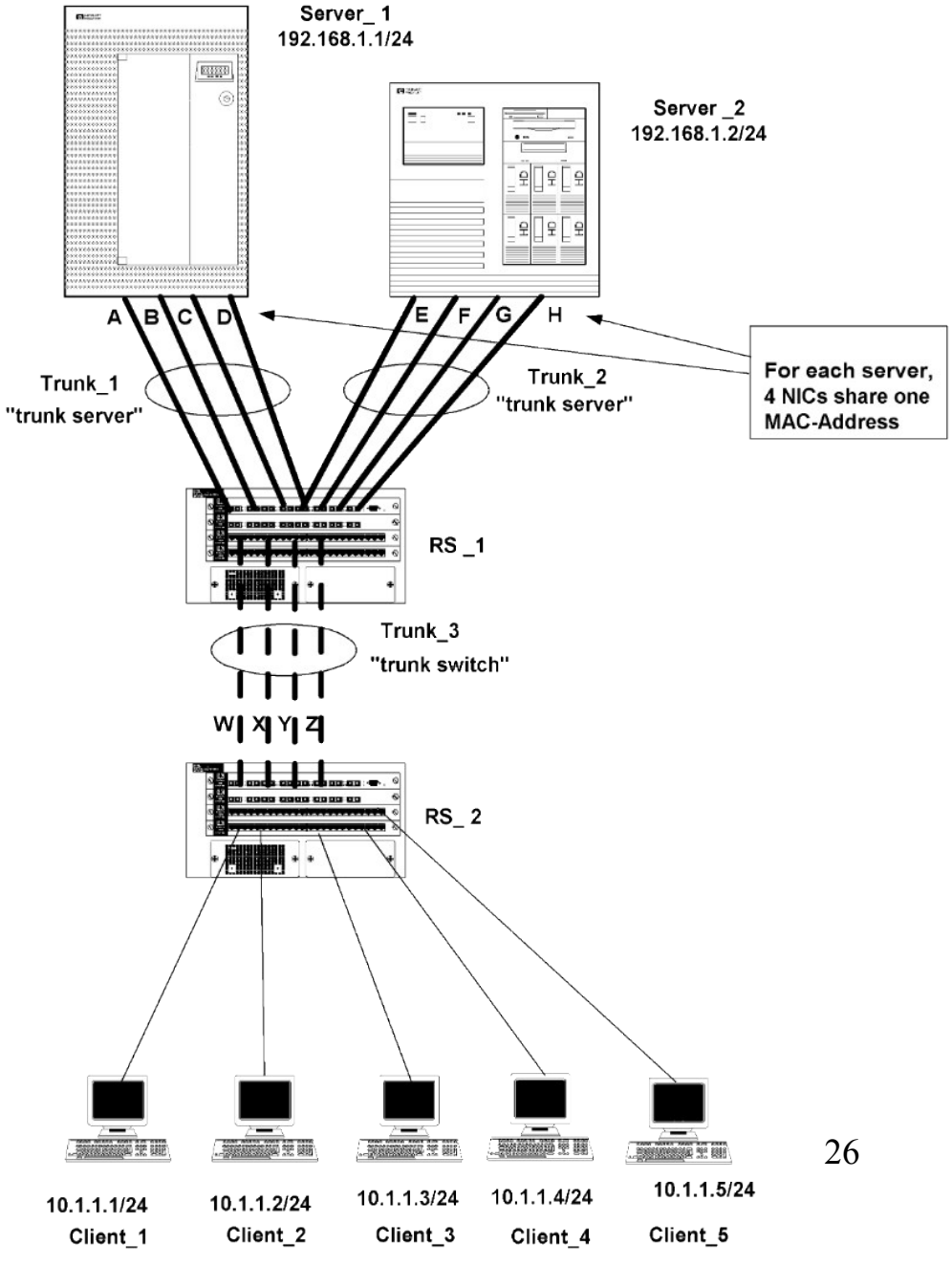
Redondance au niveau matériel

Les cartes réseaux

Problème:

Il est important que les équipements supportent le 802.3ad car la même adresse MAC est partagée entre les cartes réseaux

En effet, cela entraînerait une confusion pour l'équipement qui mettrait à jour ces tables ARP en permanence et introduirait ainsi de la latence !

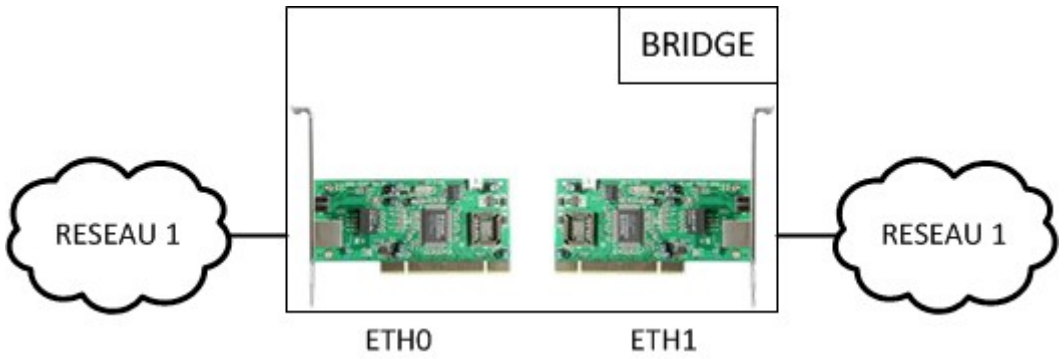


Les cartes réseaux

Solution:

Si la problématique est la tolérance de panne, il est possible d'utiliser le mode bridge de Linux pour "agréger" plusieurs cartes.

(On n'oubliera pas d'activer le Spanning Tree...)



Les serveurs lames / blades

Les principaux avantages des châssis lames sont

- un coût acquisition et d'exploitation inférieur ;
- un délai de mise en service inférieur ;
- dans le cas de serveurs ayant un accès SAN on peut rendre les lames interchangeables entre elles (Boot On San).
- une consommation électrique réduite (mutualisation) ;
- un dégagement thermique réduit ;
- un encombrement réduit (~ 60 serveurs / m²).

Redondance au niveau réseau

Redondance du backbone (OSI 2) :

Les réseaux ethernet doivent fournir un chemin entre deux points (topologie en étoile).

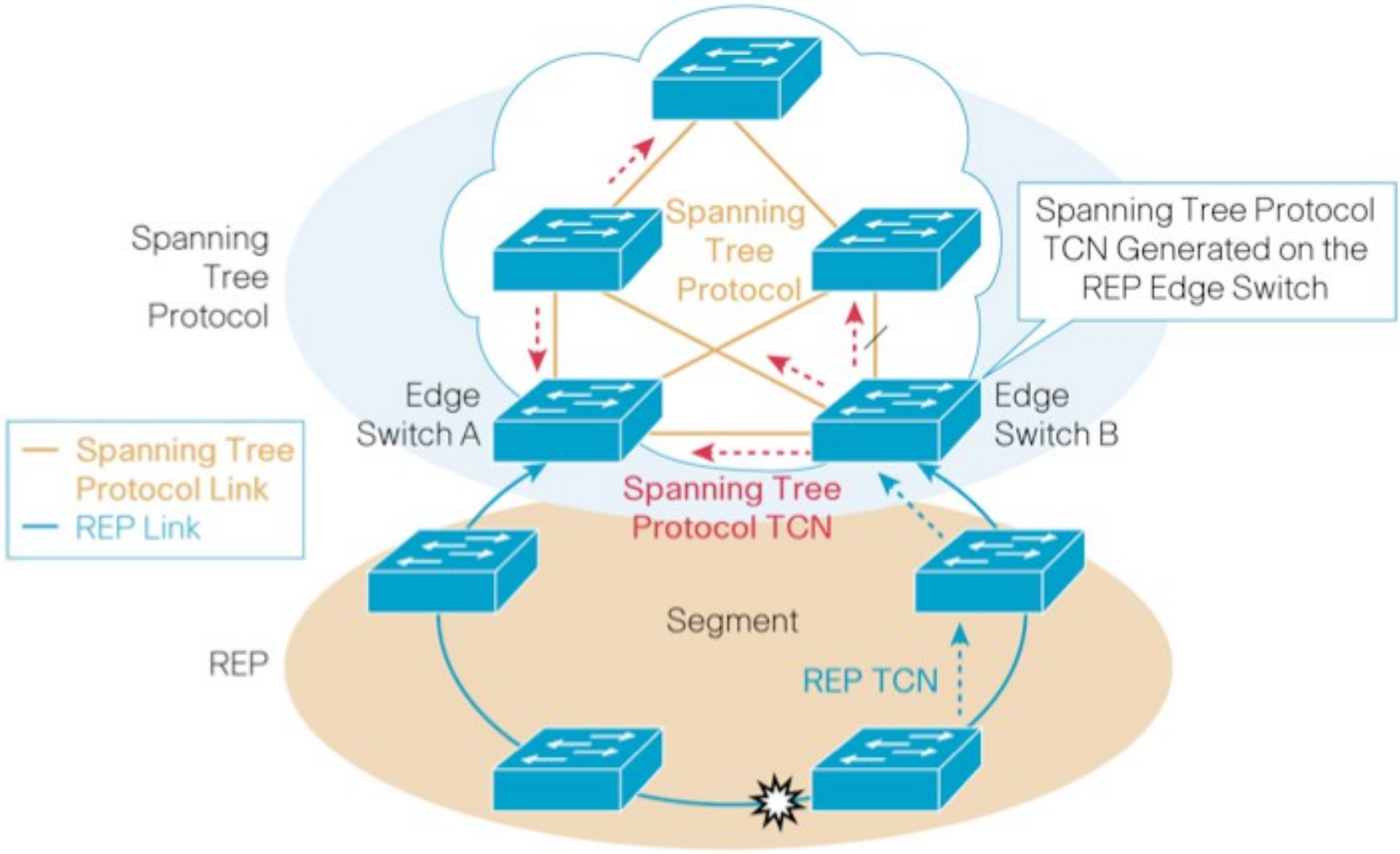
Problème → si la liaison tombe, la connexion est coupée (SpoF) ...

Solution → faire une boucle entre ces deux points...

Problème → tempête de broadcast...

Solution → utiliser le Spanning Tree !

Redondance du backbone (OSI 2) :



Couverture du sans fil (OSI 2) :

Les réseaux sans-fil doivent permettre de se déplacer sans contrainte.

Problème → si le périphérique sort de la zone de couverture, la connexion est coupée...

Solution → utiliser plusieurs points d'accès répartis "habilement"...

Problème → ces points d'accès doivent avoir le même SSID et pouvoir "se passer" les périphériques sans-fil...

Solution → utiliser 802.11r (roaming) !

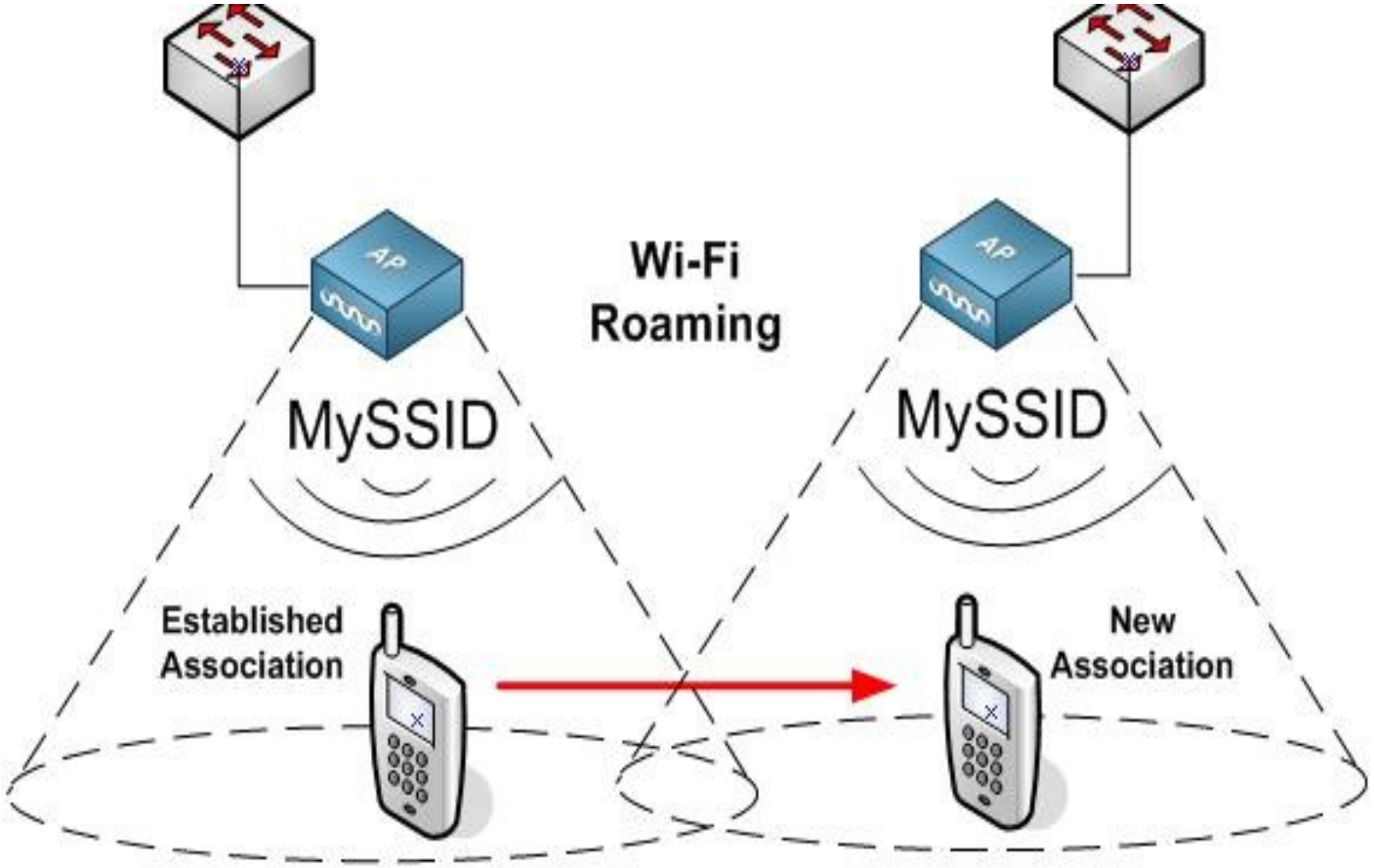
Couverture du sans fil (OSI 2) :

Sans 802.11r, la transition passe par six étapes :

1) balayage (identification des AP) ;
2) authentification 802.11 avec le nouvel AP ;
3) réassociation (connexion avec le nouvel AP) ;
4) négociation de la clé maître ;
5) dérivation de la clé maître en clés de session ;
6) contrôle de QoS.

Avec 802.11r les étapes 2, 3, 4 et 5 ne font plus qu'une !

Couverture du sans fil (OSI 2) :



Redondance des points de sortie (OSI 3) :

Le réseau local possède des points de sortie qui permettent aux machines d'accèder à d'autres réseaux.

Problème → si la liaison tombe (ADSL, fibre, etc...), les machines sont "coupées du monde" (SPoF)...

Solution → ajouter une deuxième ligne ADSL ou fibre...

Problème → comment paramètrer l'équipement pour utiliser la deuxième connexion "au cas où"...

Solution → utiliser des routes statiques flottantes ;

Redondance des points de sortie (OSI 3) :

Les routes statiques sont utilisées lorsqu'il n'existe aucune route dynamique vers la destination, ou lorsqu'on ne peut pas exécuter un protocole de routage dynamique.

Par défaut, les routes statiques ont une distance administrative égale à **un**, qui leur donne la **priorité** sur les routes des protocoles de routage dynamique.

Lorsqu'on augmente la distance administrative à une valeur supérieure à celle d'un protocole de routage dynamique, la route statique devient un filet de sécurité en cas d'échec du routage dynamique.

Redondance des points de sortie (OSI 3) :

Exemple :

Les routes dérivées du protocole IGRP (Interior Gateway Routing Protocol) ont une distance administrative par défaut de 100.

Afin de configurer une route statique qui est remplacée par une route IGRP, il faut spécifier une distance administrative supérieure à 100 pour la route statique.

Ce genre de route statique est qualifiée de « **flottante** » car elle est installée dans la table de routage seulement quand la route préférée disparaît.

Application : ip route 172.31.10.0 255.255.255.0 10.10.10.2 101.

Redondance des points de sortie (OSI 3) :

La liaison de sortie n'est pas le seul problème, l'équipement lui-même peut défaillir.

Problème → si l'équipement tombe, les machines ne peuvent plus accéder à d'autre réseaux (passerelle injoignable)...

Solution → mettre une deuxième passerelle...

Problème → il faut "brancher" la passerelle de backup sur le lien ADSL (1 IP publique) et reconfigurer tous les ordinateurs du LAN pour utiliser la nouvelle passerelle (gateway)...

Solution → utiliser des routeurs virtuels

Redondance des points de sortie (OSI 3) :

Plusieurs implémentations de routeurs virtuels sont disponibles :

- Hot Swap Redundancy Protocol (HSRP → Cisco)
- Virtual Router Redundancy Protocol (VRRP → standard RFC 5798)

Redondance des points de sortie (OSI 3) :

Principe de fonctionnement :

Un "groupe de redondance" se voit attribuer une adresse IP partagée entre les membres du groupe.

Au sein de ce groupe, un hôte est désigné comme "maître".

Les autres membres sont appelés "esclaves".

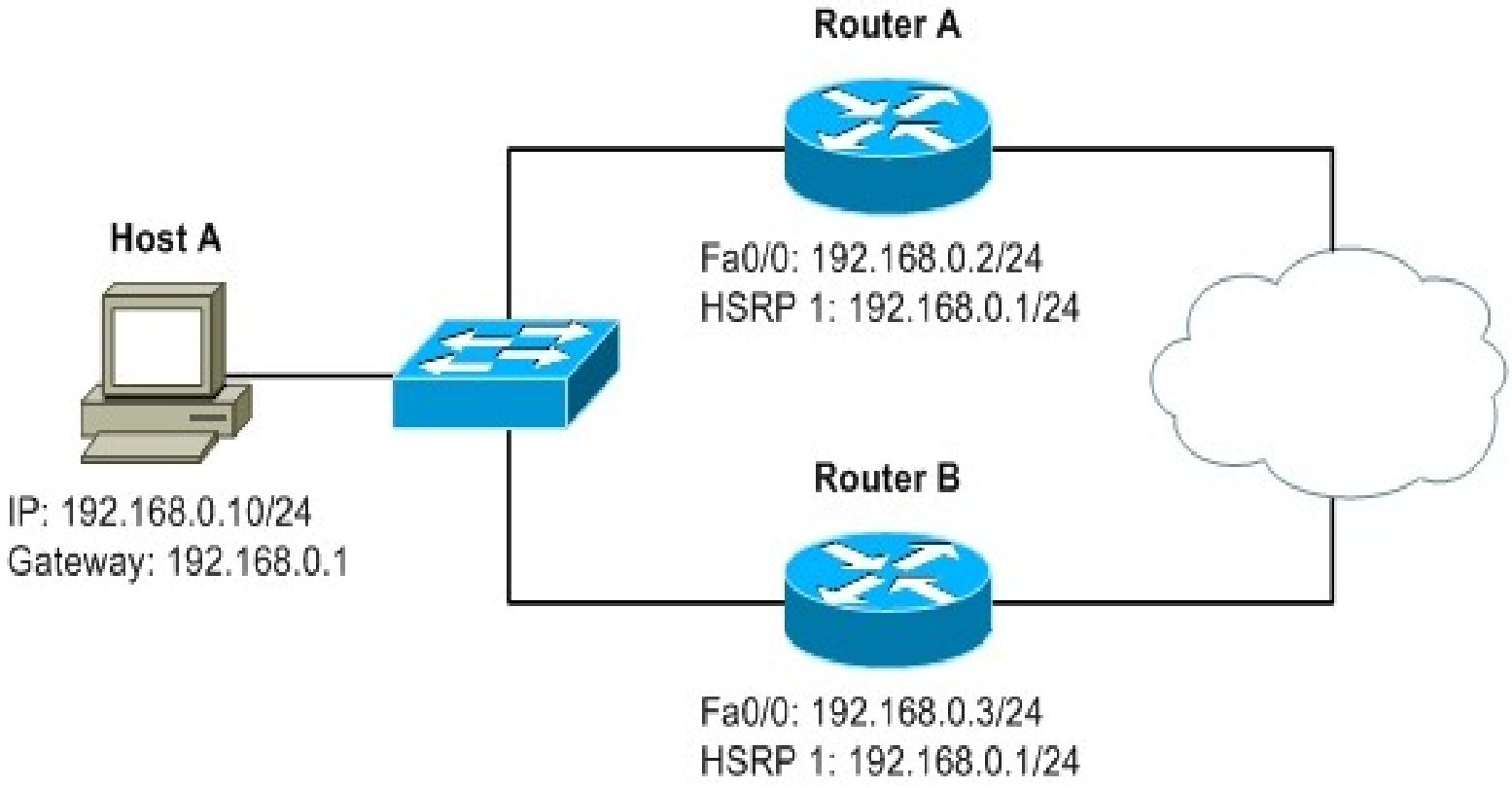
L'hôte maître est celui qui "prend" l'adresse IP partagée.

Il répond à tout trafic ou requête ARP à l'attention de cette adresse.

Chaque hôte doit avoir une seconde adresse IP unique.

Redondance des points de sortie (OSI 3) :

Principe de fonctionnement :



Redondance au niveau service

Système de maître / esclave :

Un périphérique, un processus ou un serveur est le maître, le ou les autres sont les esclaves.

Le maître donne des ordres a l'esclave qui les exécute et possède les données de références qui sont "poussées" sur les esclaves

Ces échanges peuvent être sécurisés (eg. RNDC)

L'avantage d'un tel système est :

- de faire la configuration du maître et qu'elle soit répliquée à l'identique sur tous les esclaves ;
- d'offrir une tolérance à la panne.

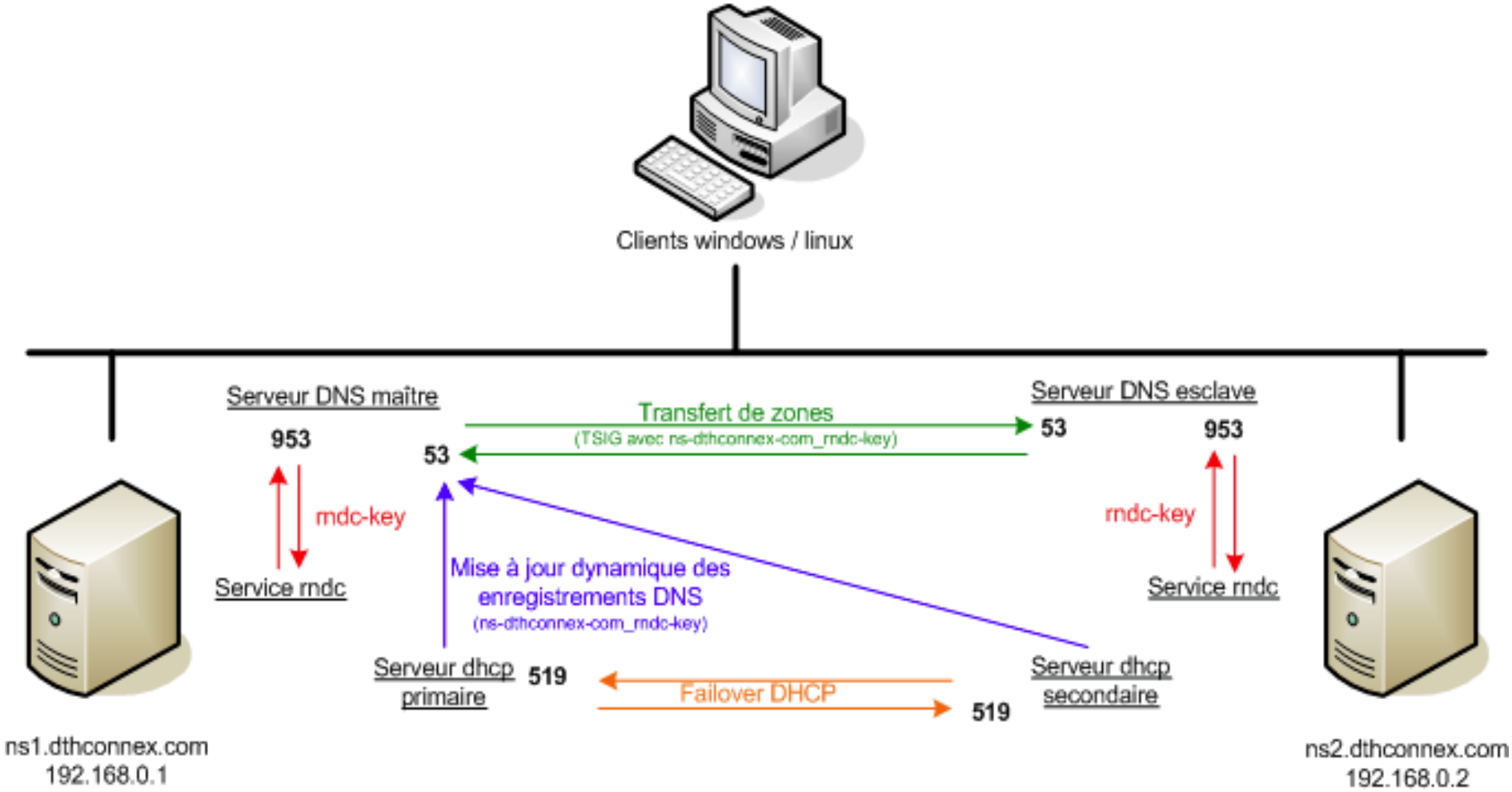
Système de maître / esclave :

Beaucoup de services offrent de tels mécanismes :

- BIND (avec RNDC) ;
- DHCPd ;
- Samba (avec LDAPs → utilise SSL) ;
- Postfix
- ...

Système de maître / esclave :

Exemple de BIND avec Remote Name Daemon Control (RNDC)



Système de partage d'IP :

C'est le même mécanisme que pour la partie réseau:

- Un maître possède l'IP de référence ;
- Les esclaves possèdent tous une IP unique ;
- La bascule s'opère quand le maître ne répond plus aux paquets "HELLO"

Les implémentations sont :

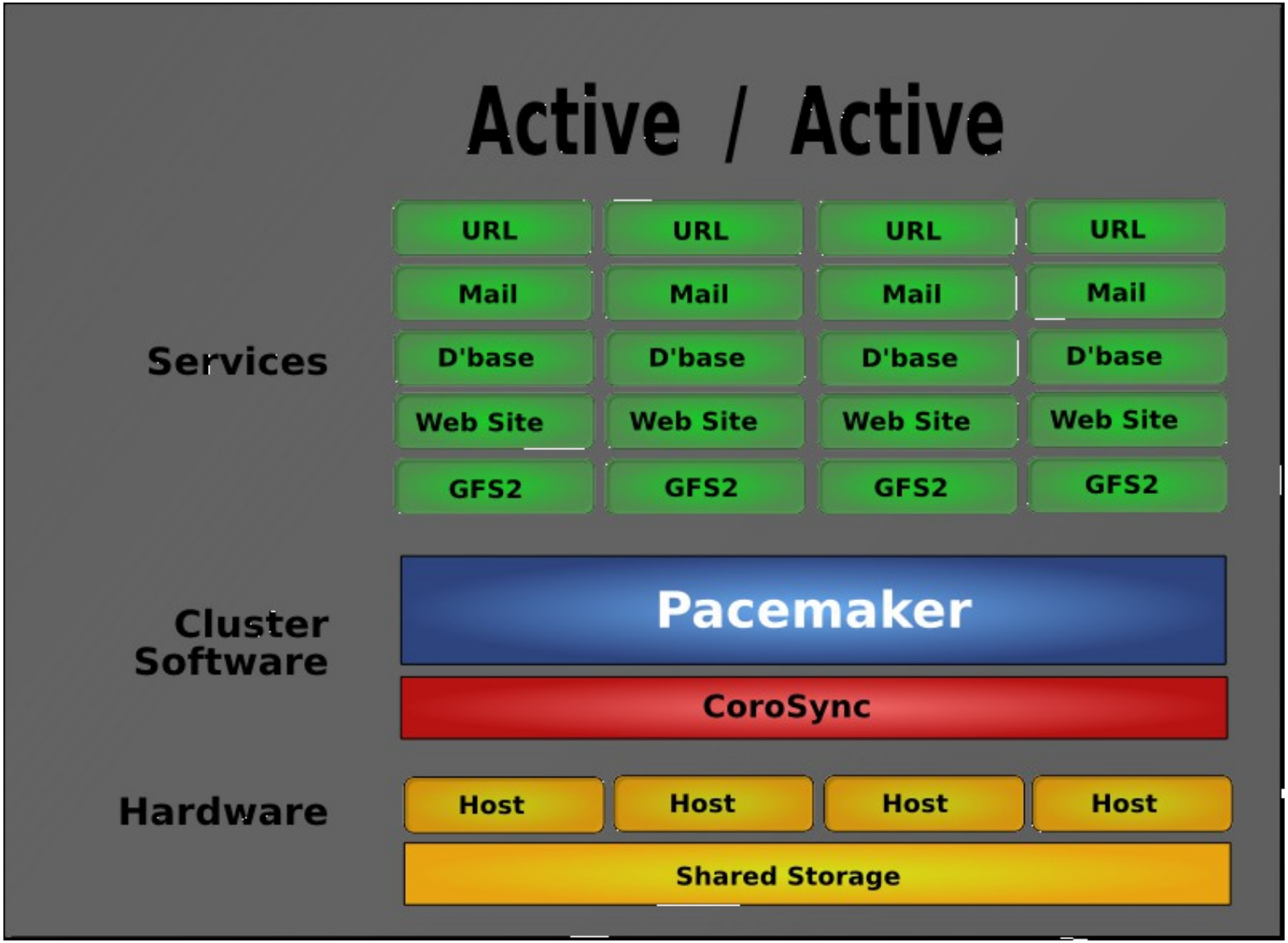
- CARP (BSD) / uCARP (Linux)
- HeartBeat (HA Project)
- PaceMaker / Corosync

Système de partage d'IP :

HeartBeat et le couple PaceMaker / Corosync permettent de faire plus que du partage d'IP et notamment :

- du clustering ;
- un fonctionnement actif / actif ;
- de la redondance au niveau système.

Système de partage d'IP :



Système de synchronisation à chaud :

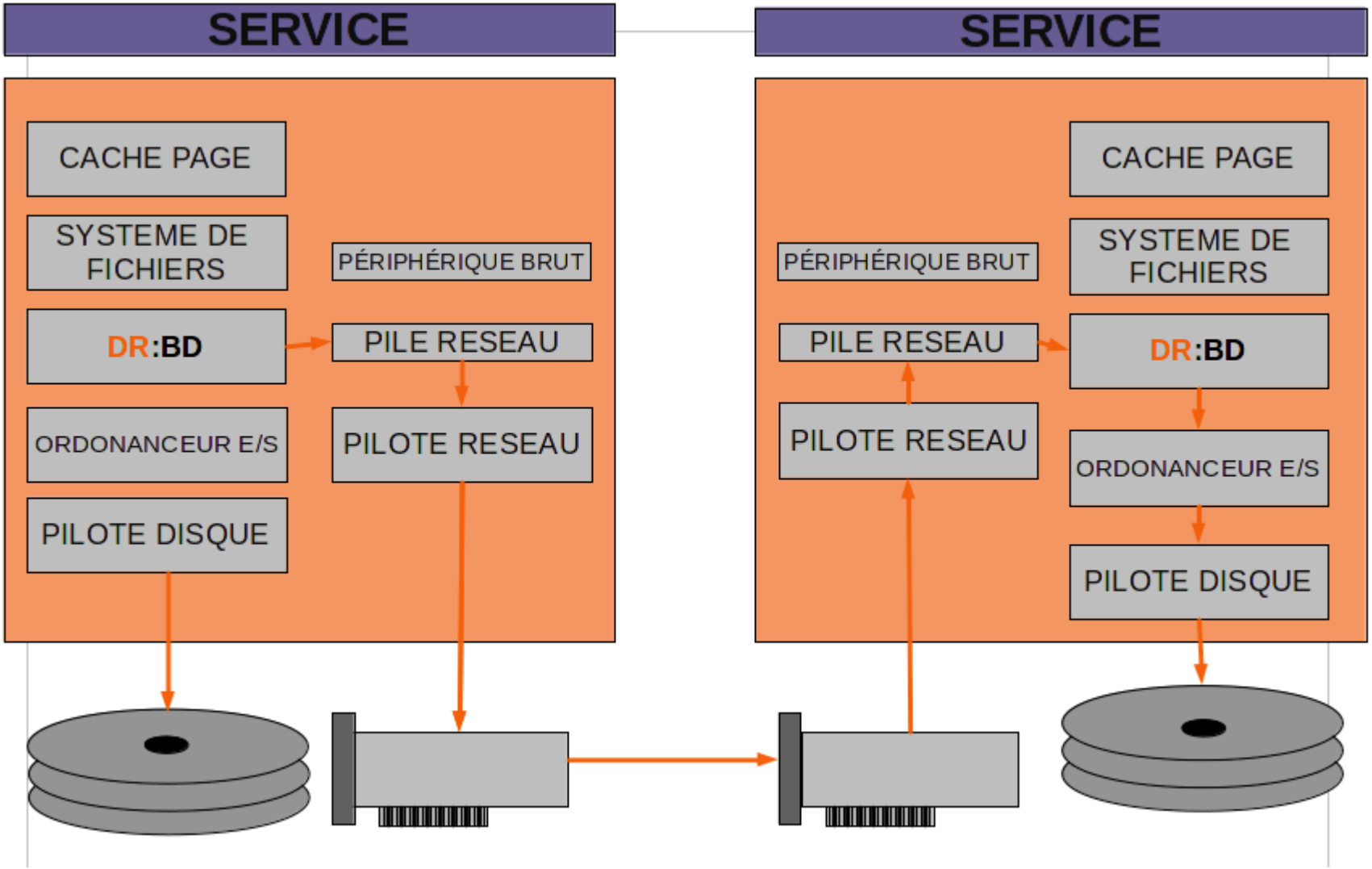
DRBD ajoute une couche logique de périphériques de blocs au-dessus de la couche logique locale des périphériques de blocs

Les écritures sur le nœud primaire sont simultanément propagées au nœud secondaire.

Le nœud secondaire transfère ensuite les données à son périphérique de bloc de bas niveau correspondant.

Toutes les lectures sont effectuées localement.

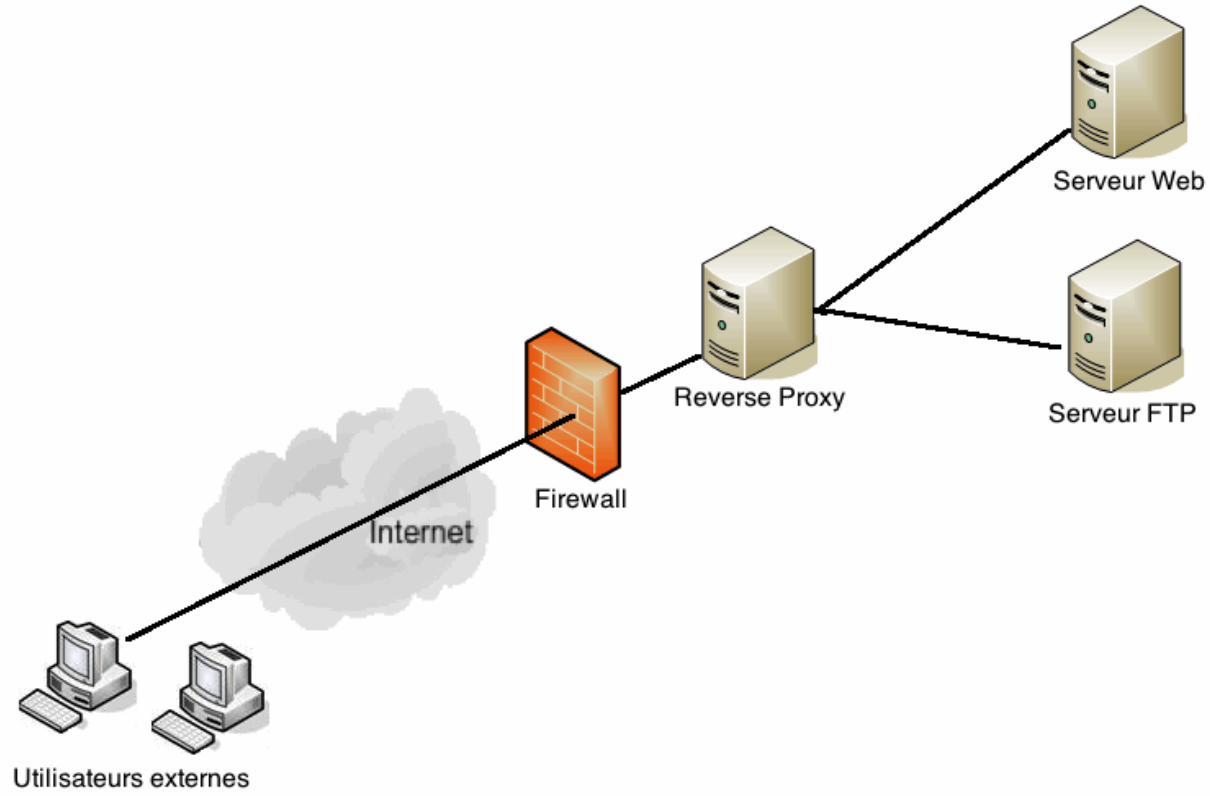
Système de synchronisation à chaud :



Reverse proxy :

Le proxy inverse est installé du côté des serveurs Internet.

L'utilisateur du Web passe par son intermédiaire pour accéder aux applications de serveurs internes.



Reverse proxy :

Les fonctionnalités offertent par le proxy inverse sont :

- une mémoire cache (contenu statique) ;
- un intermédiaire de sécurité (URL rewrite, …) ;
- le chiffrement SSL (mode bump) ;
- la répartition de charges (RR, DR, …)
- la compression.

Pour que les sessions soient conservées, les proxys inverses utilisent deux mécanismes:

- Sticky session (ajout d'un cookie dans le header) ;
- Source hash (calcul d'un hash en fonction du port + IP sources)

Balance de charge :

La balance de charge permet de répartir l'effort sur plusieurs machines rendant le même service (cluster).

Cette balance peut se faire de plusieurs façons :

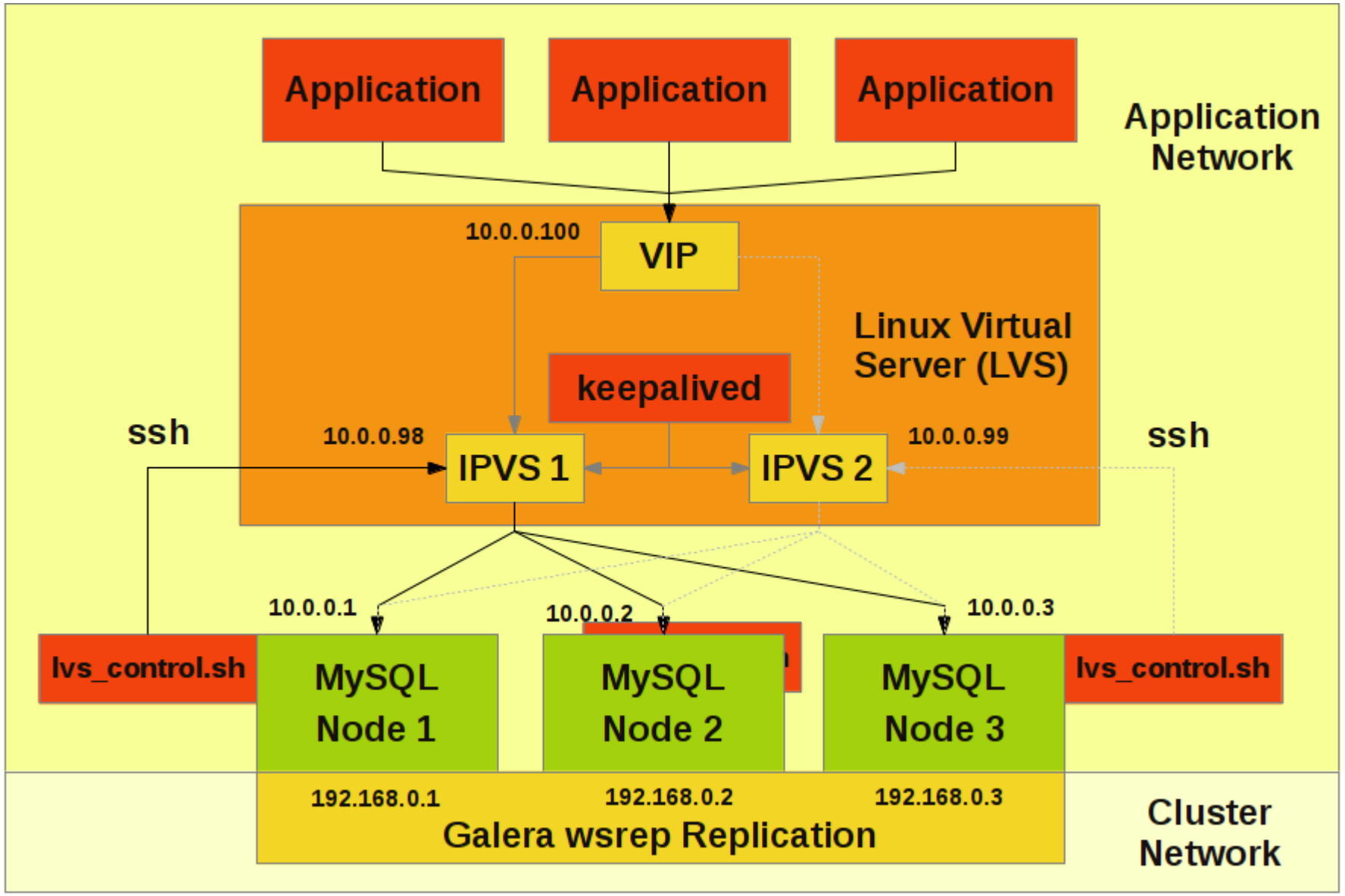
- RR (Round-Robin) ;
- Least-Connection ;
- Destination / Source Hashing ;
- Shortest Expected Delay.

Balance de charge :

Les protagonistes sont les suivants :

- LVS director : c'est le répartiteur de charge qui reçoit toutes les connections clientes et qui les redirige vers un "vrai serveur" ;
- "vrai" serveur : élément du noeud qui forme le cluster LVS et qui fournit le service ;
- clients : ordinateurs faisant des requêtes sur le serveur virtuel (cluster).

Balance de charge :



Redondance au niveau datacenter

Image d'exploitation :

La virtualisation permet de faire des snapshots (images de machines) virtuels.

Ces snapshots permettent de capturer l'état entier de la machine au moment où ils sont déclenchés.

Ils peuvent être utilisés pour faire un point de contrôle d'un système d'exploitation, sauvant ainsi d'une erreur de configuration.

Copie à chaud :

La copie à chaud permet de sauvegarder la machine virtuelle complète sur un serveur distant, pendant son fonctionnement (aucun downtime)

Gestion de configuration :

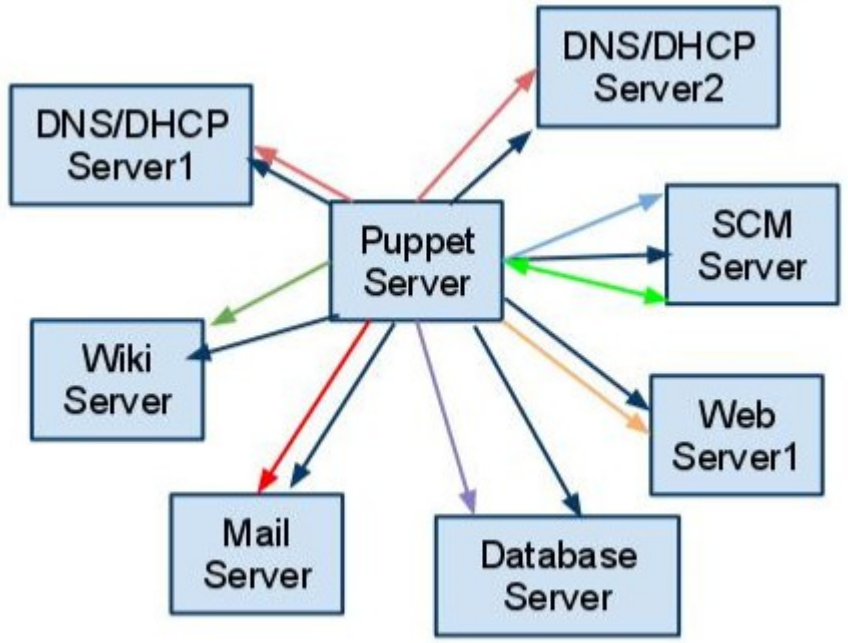
La gestion de configuration consiste à gérer la description technique d'un système

Elle permet la gestion de systèmes complexes ainsi que leurs déploiements.

Gestion de configuration :

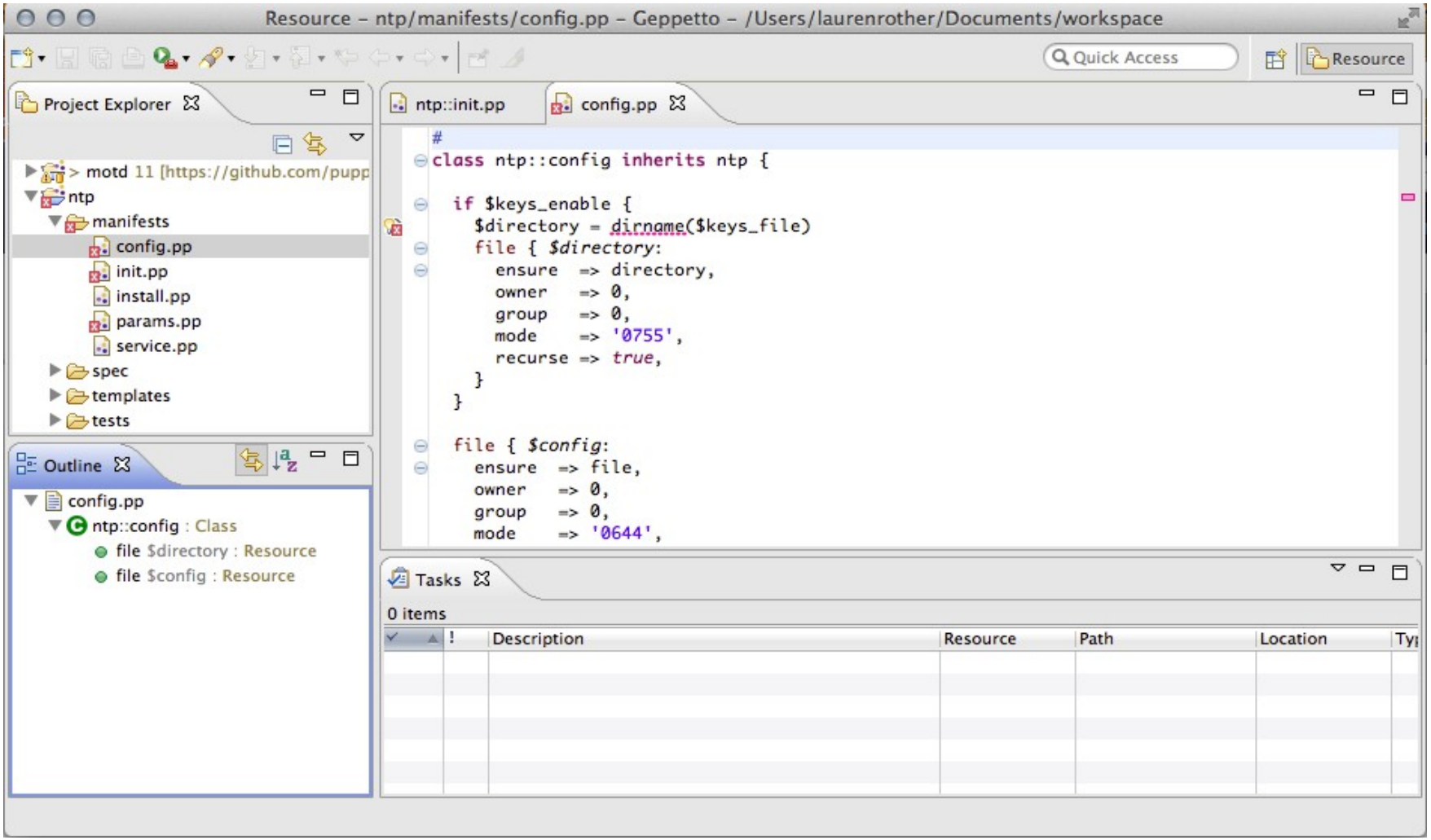
Un des outils les plus utilisés est Puppet.

Il permet de décrire les configurations dans un langage déclaratif qui forme des scripts de déploiement (catalogue)



Gestion de configuration :

Il existe un pluggin Eclipse qui permet d'écrire ces catalogues (Geppeto)



Constitution de noeuds :

Les serveurs de virtualisation n'échappent pas à la régle et sont eux-mêmes sujets aux pannes.

Le cluster ou noeud comprend plusieurs serveurs de virtualisation qui sont dédiés au fonctionnement des **mêmes** machines virtuelles.

Cela permet, à l'image du cloud, de dématérialiser l'endroit où la machine virtuelle s'exécute.

De la sorte, la montée en charge peut être anticipée et la tolérance aux pannes est plus grande.

Constitution de noeuds :

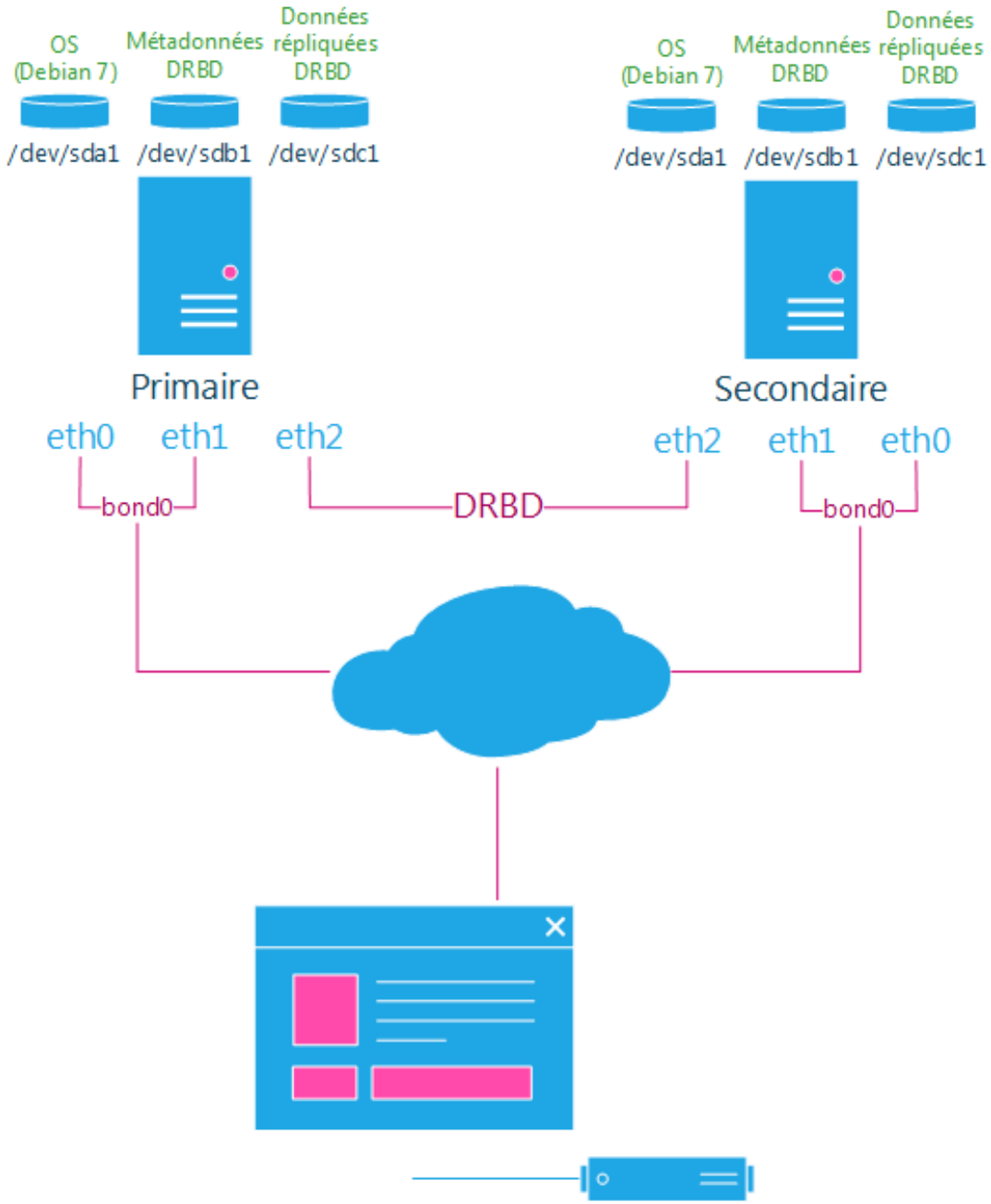
Les avantages du noeud avec **Proxmox** sont :

- Interface web de gestion centralisée ;
- Support de plusieurs sources d'authentification (eg. local, MS ADS, LDAP, …) ;
- Gestion des ressources par rôle (VM, stockage, noeuds, etc.) ;
- Clusters multi-maîtres ;
- Migration de VM entre noeuds physiques.

Redondance au niveau datacenter

Constitution de noeuds :

Exemple de déploiement



Engagement utilisateur / service

La disponibilité d'une application est inversement proportionnelle à son temps d'immobilisation.

Le temps d'immobilisation total est simplement la somme des durées de chaque panne ou interruption.

Par conséquent, pour augmenter la disponibilité d'un système, il faut diminuer la durée des pannes, réduire leur fréquence, ou les deux.

Si la panne est grave, il faudra construire un nouveau serveur à partir de zéro et restaurer toutes les données et services.

Imaginons que l'on promette un Service Level Agrement (SLA) de 99,5%, en comptant sur une seule panne par mois.

Dans les 3 heures et 43 minutes qui suivent le début de la panne, on doit dérouler les cinq phases de restauration qui sont :

1) Phase de diagnostic → diagnostiquer le problème et déterminer l'action appropriée.

2) Phase d'approvisionnement → recenser, trouver, transporter et assembler physiquement le matériel, le logiciel et le média de sauvegarde de remplacement.

3) Phase de mise en place → configurer le matériel du système et installer un OS de base.

4) Phase de restauration → restaurer tout le système à partir du média, y compris les fichiers système et les données utilisateurs.

5) Phase de vérification → vérifier le bon fonctionnement de tout le système et l'intégrité des données utilisateurs.

Indépendamment du SLA, on doit connaître la durée de chaque phase.

De plus, chaque phase peut introduire des retards inattendus !

Par exemple, une phase de diagnostic peut accaparer beaucoup de temps et il est raisonnable de passer à la phase d'approvisionnement si l'anomalie n'est pas trouvée en 15 minutes.

La phase d'approvisionnement peut elle aussi prendre du temps si le média de sauvegarde est hors site et s'il faut attendre sa livraison.

Si le camion chargé de livrer une bande de sauvegarde hors site a un accident en route...

Peut-être que 3 heures et 43 minutes est une durée bien courte pour restaurer un service, mais en réalité, on peut très bien ne disposer que de 2 heures !

C'est pourquoi il faut toujours un plan d'action et surtout, faire des "répétitions générales" pour vérifier qu'aucun grain de sable ne viendra enrailler le mécanisme de restauration du service

Tutoriaux

Raid :

https://www.tala-informatique.fr/wiki/index.php/Gestion_des_disques#mdadm

DRDB :

https://www.tala-informatique.fr/wiki/index.php/Drbd

Corosync :

http://www.unixmen.com/install-corosync-pacemaker-centos-6-5/

Puppet (exemple bind) :

https://forge.puppetlabs.com/thias/bind

Geppetto :

https://docs.puppetlabs.com/geppetto/latest/index.html

Proxomox (node) :

https://aresu.dsi.cnrs.fr/spip.php?article198